DOCUMENT RESUME

ED 346 142 TM 018 466

AUTHOR Sadek, Ramses F.; Huberty, Carl J.
TITLE On Outlier Influence in Classification Analysis.
PUB DATE Apr 92
NOTE 22p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Change; *Classification; *Computer Simulation; *Estimation (Mathematics); Groups; *Mathematical Models; Monte Carlo Methods; Sample Size
IDENTIFIERS Change Analysis; *Misclassification (Statistics); *Outliers

ABSTRACT

Using computer simulation data, the effect of a single global outlier in two-group classification analysis was explored in terms of the outcome variables of change in classification results (PCHNG), change in misclassification rate (MISDIF), and change in precision of misclassification rate estimation. The precision of misclassification rate estimation was assessed in terms of the standard deviation of misclassification rate estimates. The change of interest pertained to the standard deviation without an outlier and the standard deviation with an outlying observation. Four factors were considered: (1) number of outcome variables; (2) group size; (3) distance between the groups; and (4) location of the outlier. Results indicate that the presence of a global outlier increases the misclassification rate and influences the classification results. In general, the influence of a global outlier (in terms of MISDIF and PCHNG) increased as the sample size decreased. However, the presence of a global outlier did not affect the precision of misclassification rate estimation. Nine tables and three bar graphs present results of the Monte Carlo studies. There is a 12-item list of references. (Author/SLD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

RAMSES F. SADEK

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED346142

# *ON OUTLIER INFLUENCE IN CLASSIFICATION ANALYSIS*

RAMSES F SADEK CARL J HUBERTY

University of Georgia

329 Aderhold Hall

University of Georgia

Athens, Ga. 30602

Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1992

TM018466



BEST COPY AVAILABLE

## *ABSTRACT*

Using computer simulated data, this study has explored the effect of a single global outlier in two-group classification analysis. The effect was measured in terms of three outcome variables: (1) change in classification results (PCHNG); (2) change in misclassification rate (MISDIF); and (3) change in precision of misclassification rate estimation. The precision of misclassification rate estimation is assessed in terms of the standard deviation of misclassification rate estimates; the change of interest pertains to the standard deviation without an outlier and the standard deviation of that with an outlying observation. Four factors were considered: (1) number of outcome variables; (2) group size; (3) distance between the groups; and (4) location of the outlier.

The results indicated that the presence of a global outlier increased the misclassification rate and influenced the classification results. In general, the influence of a global outlier (in terms of MISDIF and PCHNG) increased as the sample size decreased. However, the presence of a global outlier did not affect the precision of misclassification rate estimation.


## *ON OUTLIER INFLUENCE IN CLASSIFICATION ANALYSIS*

A persisting problem common to many applied disciplines is that of assigning an object (observation) to one of two or more distinct groups on the basis of the similarity of that object's scores on a set of variables to group profiles composed of the same set of variables. Classification is one name that is often attached to this problem.

A large number of methods of classification analysis has been proposed. They are surveyed in Hand (1981). The earliest classification analysis method to be developed is based on the two-group linear discriminant function method of R. A. Fisher. Fisher's procedure is equivalent to the likelihood ratio procedure that arises under the assumption of multivariate normality with a common covariance matrix (Panel on Discriminant analysis, Classification, and Clustering, 1989, p. 36). The linear classification rule has the property that it is optimal for the normal class equal-covariance matrix situation, as well as for some categorical variable situations.

## 1. OUTLIERS AND CLASSIFICATION ANALYSIS

Hawkins (1980, p. 1) defined an outlier as an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism. Barnett and Lewis (1984, p. 4) defined an outlier in a set of data to be an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data.

Anscombe (1960) listed three causes of outliers:

(1) Inherent variability, the variability of stature that would be observed in the population even if all measurements were perfectly accurate. This variability cannot be reduced without changing the population itself, the object of the study.

(2) Measurement error, the error in using the measuring instruments.

(3) Execution error, a name we may use to include any discrepancy between what we intend to do and what is actually done, other than error in the use of measuring instruments.

Beckman and Cook (1983) listed five reasons for studying outliers: (1) Special interest in the discordant observation itself rather than in, for example, the estimation of a population parameter; (2) Detection of a specific alternative or rare phenomena rather than of estimating characteristics of common, known phenomena; (3) Diagnostic indicators, where the presence of outliers is often an indication of weaknesses in the model, the data, or both; (4) The study of the nature and frequency of outliers in a particular problem can lead to models and estimation methods that accommodate outliers and thus result in improved inferences; and (5) Studying the

influence of some outlying observations to obtain a better understanding of the structure of a problem; for example, such studies may uncover inherent deficiencies in the data and lead to further experimentation.

Many procedures have been suggested for detecting outliers in multivariate data and have been reviewed by Gnanadesikan and Kettenring (1972) and Barnett and Lewis (1984). Some of the techniques used for detecting multivariate outliers can be used for detecting outliers in classification analysis. However, detecting outliers in classification analysis is more complicated than in any other analyses since the impact of an outlier on both group means and covariance matrices must be evaluated. In a two-group classification analysis, a distinction of two types of outlier may be made. A local outlier, where the observation is considered an outlier for one group, and a global outlier, where the observation is considered an outlier for both groups. Only results concerning the influence of a global outlier are given in this paper.

Due to their impact, a great deal of literature has been published on identification, deletion, and influence of outliers (see reviews by Barnett & Lewis (1984), Hawkins (1980), and Beckman & Cook (1983) for lists of references). However, no publication on the subject concerning influential observations in classification analysis has been located.

## 2. THE MONTE CARLO STUDY

This study was designed to investigate the influence of a single global outlier on two-group classification analysis results. The two groups of response variable scores were assumed to represent multivariate normal populations with a common covariance matrix. In this situation, the linear classification function rule or the minimum distance rule is appropriate.

Fifty four data conditions representing combinations of four factors were considered. The four factors are:

(1) the number of response variables, NVAR=p, with three levels of 2, 4, and 6;
(2) the group size, SIZE ($n_1 = n_2$), as a multiple of p, with three levels of 3p, 6p, and 9p;
(3) the degree of separation between the two groups, in terms of Mahalanobis distance, DISTANCE= $\Delta$, with values of .5, 1, and 1.5; and
(4) the location (LOC) or extremeness of the outlier, in terms of a P-value associated with an F test (see the Appendix) used for outlier detection based on the sample Mahalanobis distance, with two ranges of P-values, namely, a P-value below .02 (LOC=.02) and a P-value between .02 and .05 (LOC=.05).

For each of the 54 data conditions, two data sets were generated from multivariate normal distributions to represent the two groups. The identity matrix represented the common covariance matrix. No loss of generality was sacrificed by using $\Sigma = I$, because the discriminant function can be thought of as first reducing the covariance matrix to its principal components and then standardizing so that the variation is equal in all directions. A similar argument is given by Hand (1981, p. 139). The parameters representing the mean vectors (p elements) for the data sets from the first and second group were: $\mu_1^T = (2,\ldots,2)$, and $\mu_2^T = \mu_1^T - (\Delta/\sqrt{p})(1,\ldots,1)$, respectively. SAS/IML (version 6.07 on a UNIX system) was used to generate the data sets.

After generating each pair of data sets, the linear classification rule was used to classify each observation into one of the two groups. The misclassification rate was calculated and recorded. A single global outlier was inserted at the end of the second group. A check was made to confirm that the inserted observation was in fact a global outlier (using the exact distribution of the sample squared Mahalanobis distance given in the Appendix). The linear classification rule was rebuilt, using all data points including the outlier, and each observation was again classified to

one of the two groups. The classification results were recorded as before. The entire process was repeated 2000 times for each of the 54 data conditions, resulting in 108,000 replications. The apparent error rate (resubstitution method) was used.

For each of the 108,000 replications, the values of the following variables were calculated

PCHNG : the percent of the observations for which classification results were affected by the presence of the outlier; i.e., the proportion of those observations that were correctly classified in the absence of the outlier, but were misclassified in the presence of the global outlier, and those observations that were misclassified in the absence of the outlier and correctly classified in the presence of the outlier. The following box is used to explain the idea.

| | | Outlier absent 1 | 0 | |
|---|---|---|---|---|
| Outlier present | 1 | a | b | n1. |
| | 0 | c | d | n2. |
| | | n.1 | n.2 | N |

1 : correctly classified; 0: misclassified

PCHNG= (b+c)/N; where N is the total number of observations in the data set excluding the outlier.

MISWO: The misclassification rate in the absence of the outlier; MISWO= (n.2/N).

MISWITH: The misclassification rate in the presence of outlier. MISWITH= number of misclassified observations in the data (including the outlier), divided by the total number of observations (N+1).

MISDIF: Change of misclassification rate, MISWITH-MISWO.

A four-way analysis of variance on each of PCHNG and MISDIF was conducted. The four factors involved were the distance between the groups (DISTANCE), the number of response variables (NVAR), the size of the data set as a multiple of number of variables (SIZE), and the location of the outlier in terms of P-value associated with an F-statistic (LOC). SAS (version 6.07 on a UNIX system) GLM procedure was used for the analyses of variance, and both SAS MEANS and SAS FREQ procedures were used to obtain the descriptive statistics.

Partial $\eta^2$ values are used along with P-values associated with F statistics to determine the "significance" of the effects. A discussion of a cut-off point for $\eta^2$ is given by Keppel (1991, pp. 66-68). The cut-off points of .05 for $\eta^2$ along with .01 for P-value are employed in the current study. A result is considered "significant" if and only if $\eta^2$>.05 & P<.01. The research questions of interest are as follows:

1. Does the presence of a single global outlier affect the classification results in terms of PCHNG?
2. Does the presence of a single global outlier change the misclassification rate? Is MISDIF significantly different from 0?
3. Does the presence of a single global outlier affect the precision of the misclassification rate estimation? That is, is the standard deviation of misclassification rate estimates affected by the presence of a single global outlier?

## 3. RESULTS

In this section, results are given concerning the three potential influences of a single global outlier as indicated by the above listed three research questions for two-group classification results. The results concerning a local outlier are available upon request from the authors. It may be noted that the total sample size (N) is directly related to NVAR and SIZE through the relation: N = 2 x SIZE for a given value of NVAR (p),

where the multiplier 2 represents the number of groups. For example, N=48 for SIZE= 6p, and NVAR= 4.

3.1. PCHNG

The PCHNG mean values ranged from .0186 (NVAR=6, SIZE=9p, DISTANCE = 1.5, LOC=P05) to .3725 (NVAR=2, SIZE=3p, DISTANCE=.5, LOC=P02) across all 54 conditions (see Table 1), with a grand mean of .1144 (based on 108,000 replications). This indicates that in some situations, especially when the total sample size ($N= n_1+n_2$) is small, more than 37% of the sample observations may change group membership.

*Insert Table 1 about here*

The analysis of variance (ANOVA) results for the PCHNG associated with a global outlier are reported in Table 2. Based on the ANOVA results, no four-way interaction effect and no three-way interaction effects were judged to be significant.

*Insert Table 2 about here*

The only significant two-way interaction was found between NVAR and SIZE, where NVAR is the number of response variables, and SIZE is the group size as a multiple of NVAR ($F(4,107946)= 1890.60$, $P=.0001$, $\eta^2= .0655$). NVAR and SIZE main effects were found significant; however, there is no need to look at them since their interaction was found significant and will be further investigated.

As the sample size increased the influence of a global outlier in PCHNG decreased (see Table 3 and Figure 1). For example, based on 12,000 replications (for each of the 9 levels of NVAR*SIZE), the mean of PCHNG decreases from .2819 (for NVAR= 2, SIZE= 3p with a total sample size of 12) to .0842 (for NVAR= 4, SIZE= 6p with a total sample size of 48) to .0500 (for NVAR= 6, SIZE= 9p with a total sample size of 108). The mean value of PCHNG for a given sample size did not change significantly when the combination of NVAR and SIZE changed, for example, the mean value of PCHNG was .1016 for NVAR= 2, SIZE= 9p (with a total sample size of 36), and the mean value of PCHNG was .1015 for NVAR= 6, SIZE= 3p (with a total sample size of 36).

*Insert Table 3 and Figure 1 about here*

To test for the significance of differences between the levels of a given effect (or a contrast), t statistics along with $\eta^2$ values are used. Table 4 gives the values of t-statistics, P-values, and $\eta^2$ for the 36 contrasts represented by the differences of levels of NVAR*SIZE. When interpreting the significance of the differences between pchng means at the levels of NVAR*SIZE, the 36 contrast tests performed will be taken into consideration; that is, a P- value for each test is obtained as follows: $P=1-(1-$ the two sided tail area$)^{36}$. All contrasts involving NVAR= 2 and SIZE= 3 (e.g., when the group size=6) are significant, as can be seen from the first column of Table 4. Other than effects involving NVAR*SIZE= 2*3p, the only significant difference is found between NVAR*SIZE=2*6p (group size= 12) and NVAR*SIZE= 6*9p (group size= 54). However, the difference between other pairs of effects with the same group sizes are not significant (e.g., NVAR*SIZE= 4*3p with a group size= 12, and NVAR*SIZE= 6*9p with a group size= 54).

*Insert Table 4 about here*

DISTANCE and LOC main effects were judged to be significant ($F(2,107946)=4092.95$, $P=.0001$, and $\eta^2=.0705$ for DISTANCE, and $F(1,107946) =11474.85$, $P=.0001$, and $\eta^2=.0961$ for LOC). Based on 54,000 replications, the mean of PCHNG was .1442 when LOC=P02 and .0854 when LOC= P05. Based on 36,000 replications, the mean values of PCHNG for the levels of DISTANCE were .1473, .1098, and .0861 for DISTANCE= .5, 1.0, and 1.5, respectively. Only the difference between mean of PCHNG at DISTANCE=0.5 and mean of PCHNG at DISTANCE=1.5 was found significant ($t(71998)= 89.716$, $P=.0000$, and $\eta^2= .0694$).

### 3.2. MISDIF

Table 5 shows that MISDIF mean values were positive and ranged from .0068 (NVAR=6, SIZE=9P, DISTANCE=.5, LOC=P05) to .1816 (NVAR=2, SIZE=3p, DISTANCE=1.5, and LOC= P02), with a grand mean of .044 (based on 108,000 replications). This suggests that the presence of a single global outlier increases the misclassification rate, especially for a small total sample size ($n_1+n_2$).

*Insert Table 5 about here*

The analysis of variance (ANOVA) results for the MISDIF associated with a global outlier are reported in Table 6. Based on the ANOVA results, no four-way interaction effect and no three-way interaction effects were judged to be significant.

*Insert Table 6 about here*

The only significant two-way interaction was found between NVAR and SIZE ($F(4,107946)=2204.8$, $P=.0001$, $\eta^2=.07555$). NVAR and SIZE main effects were interpreted to be significant, however, no need to look at them since their interaction was found significant and will be further investigated. The other two main effects (DISTANCE and LOC) were judged to be not significant.

Table 7 gives means and standard deviations of MISDIF for each combination of NVAR*SIZE. As the sample size ($N=n_1+n_2$) increased the influence of a global outlier in MISDIF decreased (see Table 7 and Figure 2). For example, based on 12,000 replications for each of the 9 levels of NVAR*SIZE, the mean of MISDIF decreases from .1437 (NVAR=2, SIZE=3p with N=12) to .025 (NVAR=4, SIZE=6p with N=48) to .0105 (NVAR= 6, SIZE=9p with N= 108).

*Insert Table 7 and Figure 2 about here*

Table 8 gives the values of t-statistics, P-values, and $\eta^2$ for the 36 contrasts represented by the differences of levels of NVAR*SIZE (each of the t-values given in Table 8 is based on 5998 degrees of freedom). Table 8 shows that all contrasts involving NVAR=2 and SIZE=3p (e.g. when the group SIZE=6) are significant, as can be seen from the first column of the table. Note that as the difference between the group sizes involved in a contrast increases, the t-value and $\eta^2$ increase. For example, $\eta^2=.1332$ for the contrast 2*3p versus 2*6p (group size 6 versus 12), and $\eta^2 = .1884$ for the contrast 2*3p versus 2*9p (group size 6 versus 18).

*Insert Table 8 about here*

### 3.3 Precision Of Misclassification Rate Estimation

The study indicated that the presence of a single global outlier did not appreciably affect the precision of misclassification rate estimation. Figure 3 shows, for a global outlier, that the standard deviation of MISWO is almost identical to the standard deviation of MISWITH for every data condition.

*Insert Figure 3 about here*

## 4. CONCLUSIONS AND DISCUSSION

### 4.1. CONCLUSIONS

1. The presence of a single global outlier increases the misclassification rate by an average (across all 54 conditions studied) of approximately .05.
2. For a given data condition, the influence of a global in terms of PCHNG is greater than in terms of MISDIF.
3. The larger the sample size, the less influence an outlier has on PCHNG and on misclassification rate.
4. The more extreme the global outlier, the more influence it has on PCHNG and misclassification rate. The influence of the location of the outlier diminishes as the sample size increases.

5. The influence of a global outlier on PCHNG and misclassification rate decreases as the distance between the groups increases.

6. With a large sample size, the influence of a single global outlier on misclassification rate is negligible.

7. Although the presence of a single global outlier affects both misclassification rate (in terms of MISDIF) and classification results (in terms of PCHNG), it does not affect the precision of misclassification rate estimation.

## 4.2. DISCUSSION

The results of this study indicate that outliers can have an influence on misclassification rate and misclassification results in a two-group classification analysis. The smaller the sample size, the greater the effect a single global outlier can have on the classification results. This confirms the advice of Beckman and Cook (1983) which implies that data should always be screened for outliers prior to an analysis of the data.

Table 9 shows the conditions under which a researcher might be concerned about the influence of a global outlier on classification results measured by PCHNG and by MISDIF. The first entries (at the left of /) are for a PCHNG and the second entries (at the right of /) are for MISDIF. The total sample size ($n_1 + n_2$) is represented by N. For example, the bold entry (1/3) indicates a mean value of PCHNG for a global outlier of at least .25, and a mean value of MISDIF for a global outlier of at least .10 but less than .15 (for a total sample size of 12, DISTANCE=0.5, and LOC=P05). Such a table might be considered by a researcher in deciding whether or not to remove the outlying observation. Therefore, for example, with an N of, say, 50 -- $n_1=n_2=25$ -- and a distant global outlier and a $\Delta=1.0$ (a 4/- entry in Table 9), we need not be concerned about the misclassification rate being affected and the percent of units being affected in classification will be "low."

The study was limited to the two-group classification with equal group sizes, equal priors, and a common covariance matrix. The influence of only a single global outlier was investigated. Adding an outlier to, or removing an outlier from, a given data set may affect the equality of the covariance matrices. If this happens, the linear classification method is no longer appropriate. In the current study, the linear classification method was used with and without the existence of an outlier.

It may be noted that, in general, the impact of a local outlier in misclassification rate and on classification results was less than the impact of a global outlier. As with a global outlier, the presence of a local outlier did not affect the precision of misclassification rate estimate. For more details on the influence of a single global or local outlier, see Sadek (1992).

## *REFERENCES*


Anscombe, F. J. (1960). Rejection of outliers. Technometrics 2, 123-147.

Barnett, V. (1978). The study of outlier. Applied Statistics,27,242-250.

Barnett, V., & Lewis, T. (1984). Outliers in statistical data (3rd ed.). New York: Wiley.

Beckman, R. J., & Cook, R. D. (1983). Outlier.....s.Technometrics, 25, 119-163.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7, 179-188.

Gnanadesikan, R., & Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics, 28, 81-124.

Hand, D. J. (1981). Discrimination and classification. New York: Wiley.

Hawkins, D. M. (1980). Identification of outliers. New York:Chapman & Hall.

Panel on Discriminant Analysis, Classification, and Clustering (1989). Discriminant analysis and Clustering. Statistical Science, 4, 34-69.

Sadek, R. F. (1992). Influence of outliers in classification analysis (Doctoral dissertation, University of Georgia).
SAS Institute, Inc. (1985). SAS User's guide: Statistics. Version 5 Edition. Cary, North Carolina: SAS Institute,Inc.
Wilks, S. S. (1962). Mathematical statistics. London: Wiley.

## *APPENDIX*

### *Exact distribution of squared sample Mahalanobis distance*

Most outlier detection methods in classification analysis involve the use of a Mahalanobis distance. Assuming that the two groups are multivariate normal with mean vectors $\mu_k$ (k=1,2), and a common covariance matrix $\Sigma$, the squared sample Mahalanobis distance between an observation ($\mathbf{x}_i$) and a centroid is given by:

$$D^2_{ik}=(\mathbf{x}_i-\bar{\mathbf{x}}_k)^T\,(\mathbf{S})^{-1}(\mathbf{x}_i-\bar{\mathbf{x}}_k)\,, \qquad (1)$$

where $\bar{\mathbf{x}}_k$ is the sample mean vector for Group k and $\mathbf{S}$ is the sample pooled covariance matrix. Gnanadesikan and Kettenring (1972, p. 113) reported the exact distribution of $D^2_{ik}$ to be $(n_k-1)^2/n_k$ times a beta variable with parameters (p/2) and $(n_k-p-1)/2$. The proof is based on a theorem by Wilks (1962, p. 562). Letting $y=n_k/\{(n_k-1)^2\,D^2_{ik}\}$, the density function of y is given by:

$$f(y)=\frac{\Gamma(\frac{p}{2}+\frac{n_k-p-1}{2})}{\Gamma(\frac{p}{2})\Gamma(\frac{n_k-p-1}{2})}\,\frac{y^{\frac{p}{2}-1}}{(1+y)^{\frac{n_k-1}{2}}}, \qquad y>0 \qquad (2)$$

Let $x=y\{(n_k-p-1)/p\}$; then, the density function of x is given by

$$f(x)=\frac{\Gamma(\frac{n_k-1}{2})}{\Gamma(\frac{p}{2})\Gamma(\frac{n_k-p-1}{2})}\,(\frac{p}{n_k-p-1})^{\frac{p}{2}}\,\frac{x^{\frac{p}{2}-1}}{(1+x\frac{p}{n_k-p-1})^{\frac{n_k-1}{2}}}. \qquad (3)$$

Now, $x=\frac{n_k-p-1}{p}\,\frac{n_k}{(n_k-1)^2}D^2_{ik}$ has an F distribution with p and $(n_k-p-1)$ degrees of freedom. Thus $D^2_{ik}$ is distributed as $c_k$ times $F(p,n_k-p-1)$, where $c_k=\frac{p(n_k-1)}{n_k(n_k-p-1)}$.

Table 1

Global outlier: Means and standard deviations of PCHNG by NVAR*SIZE*DISTANCE*LOC

| LOC=P02 | | | | | LOC=P05 | | | | |
|---|---|---|---|---|---|---|---|---|---|
| NVAR | SIZE | DIST | Mean | SD | NVAR | SIZE | DIST | Mean | SD |
| 2 | 3p | .5 | .3725 | .2083 | 2 | 3p | .5 | .3128 | .2150 |
| 2 | 6p | .5 | .2483 | .1765 | 2 | 6p | .5 | .1635 | .1511 |
| 2 | 9p | .5 | .1894 | .1526 | 2 | 9p | .5 | .1133 | .1077 |
| 2 | 3p | 1.0 | .3163 | .1857 | 2 | 3p | 1.0 | .2430 | .1841 |
| 2 | 6p | 1.0 | .1696 | .1172 | 2 | 6p | 1.0 | .0933 | .0853 |
| 2 | 9p | 1.0 | .1208 | .0863 | 2 | 9p | 1.0 | .0582 | .0466 |
| 2 | 3p | 1.5 | .2715 | .1601 | 2 | 3p | 1.5 | .1753 | .1503 |
| 2 | 6p | 1.5 | .1310 | .0870 | 2 | 6p | 1.5 | .0598 | .0526 |
| 2 | 9p | 1.5 | .0893 | .0592 | 2 | 9p | 1.5 | .0387 | .0332 |
| 4 | 3p | .5 | .1942 | .1066 | 4 | 3p | .5 | .1405 | .0955 |
| 4 | 6p | .5 | .1375 | .0734 | 4 | 6p | .5 | .0785 | .0521 |
| 4 | 9p | .5 | .1072 | .0578 | 4 | 9p | .5 | .0573 | .0344 |
| 4 | 3p | 1.0 | .1672 | .0948 | 4 | 3p | 1.0 | .1027 | .0763 |
| 4 | 6p | 1.0 | .1128 | .0620 | 4 | 6p | 1.0 | .0494 | .0351 |
| 4 | 9p | 1.0 | .0826 | .0459 | 4 | 9p | 1.0 | .0342 | .0231 |
| 4 | 3p | 1.5 | .1472 | .0897 | 4 | 3p | 1.5 | .0692 | .0611 |
| 4 | 6p | 1.5 | .0917 | .0542 | 4 | 6p | 1.5 | .0353 | .0280 |
| 4 | 9p | 1.5 | .0693 | .0402 | 4 | 9p | 1.5 | .0240 | .0180 |
| 6 | 3p | .5 | .1434 | .0795 | 6 | 3p | .5 | .0973 | .0606 |
| 6 | 6p | .5 | .1080 | .0536 | 6 | 6p | .5 | .0582 | .0319 |
| 6 | 9p | .5 | .0866 | .0422 | 6 | 9p | .5 | .0426 | .0226 |
| 6 | 3p | 1.0 | .1288 | .0734 | 6 | 3p | 1.0 | .0728 | .0503 |
| 6 | 6p | 1.0 | .0914 | .0478 | 6 | 6p | 1.0 | .0379 | .0239 |
| 6 | 9p | 1.0 | .0691 | .0366 | 6 | 9p | 1.0 | .0257 | .0165 |
| 6 | 3p | 1.5 | .1138 | .0683 | 6 | 3p | 1.5 | .0529 | .0416 |
| 6 | 6p | 1.5 | .0768 | .0441 | 6 | 6p | 1.5 | .0274 | .0194 |
| 6 | 9p | 1.5 | .0572 | .0328 | 6 | 9p | 1.5 | .0186 | .0137 |

Note: each of above values is based on 2,000 replications.

Table 2

Global outlier: Analysis of variance for PCHNG

| Source | DF | Sum of Squares | F value | P | $\eta^2$ * |
|---|---|---|---|---|---|
| NVAR | 2 | 213.0844 | 12720.55 | .0001 | .1907 |
| SIZE | 2 | 201.4313 | 12024.89 | .0001 | .1822 |
| DISTANCE | 2 | 68.5618 | 4092.95 | .0001 | .0705 |
| LOC | 1 | 96.1087 | 11474.85 | .0001 | .0961 |
| NVAR*SIZE | 4 | 63.3396 | 1890.60 | .0001 | .0655 |
| NVAR* DISTANCE | 4 | 19.0493 | 568.59 | .0001 | .0206 |
| DISTANCE* SIZE | 4 | 1.6757 | 50.02 | .0001 | .0018 |
| SIZE*LOC | 2 | 1.0473 | 62.52 | .0001 | .0012 |
| NVAR*LOC | 2 | 2.4881 | 148.53 | .0001 | .0027 |
| DISTANCE* LOC | 2 | .0326 | 1.95 | .1426 | .0000 |
| NVAR*SIZE DISTANCE | 8 | .5139 | 7.67 | .0001 | .0005 |
| NVAR* SIZE*LOC | 4 | .0532 | 1.59 | .1744 | .0001 |
| NVAR*LOC* DISTANCE | 4 | .0610 | 1.82 | .1213 | .0001 |
| DISTANCE* SIZE*LOC | 4 | 1.1826 | 35.30 | .0001 | .0013 |
| NVAR*LOC* DISTANCE *SIZE | 8 | .2872 | 4.29 | .0001 | .0003 |
| ERROR | 107946 | 904.1126 | | | |
| TOTAL | 107999 | 1573.0298 | | | |

* $\eta^2_{effect} = SS_{effect} / \{SS_{effect} + SS_{error}\}$

Table 3

Global outlier: Means and standard deviations of PCHNG for NVAR*SIZE

| Level of NVAR | Level of SIZE | MEAN | SD |
|---|---|---|---|
| 2 | 3p | .28187 | .19555 |
| 2 | 6p | .14425 | .13359 |
| 2 | 9p | .10162 | .10272 |
| 4 | 3p | .13690 | .09752 |
| 4 | 6p | .08420 | .06355 |
| 4 | 9p | .06246 | .04807 |
| 6 | 3p | .10151 | .07090 |
| 6 | 6p | .06659 | .04817 |
| 6 | 9p | .04998 | .03776 |

Note: Each of these rounded values are based on 12,000 runs.

==========

Figure 1
Global outlier: Means of PCHNG for NVAR*SIZE

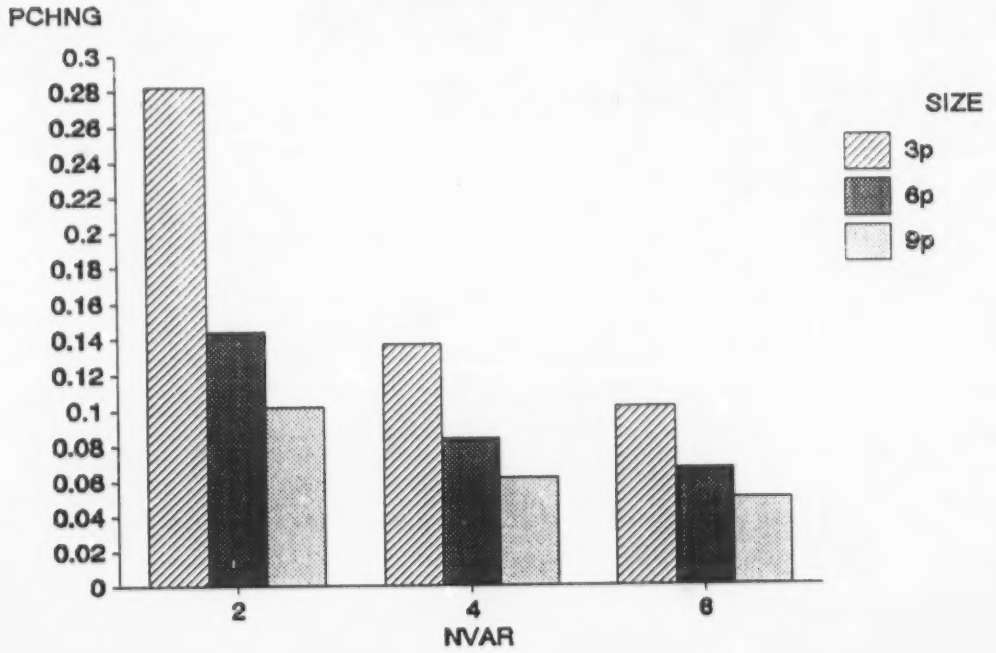

Table 4 Global outlier: t-statistic, P, and $\eta^2$ for levels of NVAR*SIZE for PCHNG

| NVAR*SIZE | 2*3p | 2*6p | 2*9p | 4*3p | 4*6p | 4*9p | 6*3 | 6*6p |
|---|---|---|---|---|---|---|---|---|
| 2*6p | 116.4770<br>.0000<br>@@ .1117 | | | | | | | |
| 2*9p | 152.5990<br>.0000<br>@@ .1775 | 36.1410<br>.0000<br>.0120 | | | | | | |
| 4*3p | 122.7230<br>.0000<br>@@ .1224 | 6.2630<br>.0000<br>.0004 | 29.8776<br>.0000<br>.0082 | | | | | |
| 4*6p | 167.3260<br>.0000<br>@@ .2060 | 50.8660<br>.0000<br>.0234 | 14.7245<br>.0000<br>.0020 | 44.1504<br>.0000<br>.0177 | | | | |
| 4*9p | 183.5760<br>.0000<br>@@ .2379 | 69.2330<br>.0000<br>.0425 | 33.0931<br>.0000<br>.0100 | 62.9693<br>.0000<br>.0354 | 18.3650<br>.0000<br>.0031 | | | |
| 6*3p | 152.6840<br>.0000<br>@@ .1776 | 36.1500<br>.0000<br>.0120 | .1093<br>1.0000<br>.0000 | 29.9614<br>.0000<br>.0082 | 14.6432<br>.0000<br>.0020 | 33.0082<br>.0000<br>.0100 | | |
| 6*6p | 182.0580<br>.0000<br>@@ .2349 | 65.7620<br>.0000<br>.0385 | 29.6227<br>.0000<br>.0080 | 59.4989<br>.0000<br>.0318 | 14.8938<br>.0000<br>.0021 | 3.4725<br>.0003<br>.0001 | 29.5380<br>.0000<br>.0080 | |
| 6*9p | 196.2720<br>.0000<br>@@ .2630 | 79.8120<br>.0000<br>@@ .0557 | 43.6719<br>.0000<br>.0174 | 73.5526<br>.0000<br>.0477 | 28.9460<br>.0000<br>.0077 | 10.5598<br>.0000<br>.0010 | 43.5884<br>.0000<br>.0173 | 14.0481<br>.0000<br>.0018 |

Note: The values in each cell are: t-statistic (based on 5998 DF), P, and $\eta^2$, respectively.

@@ indicates a significant difference.

Table 5 Global outlier: Means and standard deviations of MISDIF by NVAR*SIZE*DISTANCE*LOC

| LOC=P02 | | | | | LOC=P05 | | | | |
|---|---|---|---|---|---|---|---|---|---|
| NVAR | SIZE | DIST | Mean | SD | NVAR | SIZE | DIST | Mean | SD |
| 2 | 3p | .5 | .1500 | .1255 | 2 | 3p | .5 | .1194 | .1148 |
| 2 | 6p | .5 | .0577 | .0764 | 2 | 6p | .5 | .0375 | .0647 |
| 2 | 9p | .5 | .0337 | .0572 | 2 | 9p | .5 | .0256 | .0483 |
| 2 | 3p | 1.0 | .1715 | .1186 | 2 | 3p | 1.0 | .1220 | .1064 |
| 2 | 6p | 1.0 | .0630 | .0684 | 2 | 6p | 1.0 | .0364 | .0528 |
| 2 | 9p | 1.0 | .0356 | .0518 | 2 | 9p | 1.0 | .0227 | .0359 |
| 2 | 3p | 1.5 | .1815 | .1102 | 2 | 3p | 1.5 | .1178 | .0934 |
| 2 | 6p | 1.5 | .0682 | .0934 | 2 | 6p | 1.5 | .0389 | .0455 |
| 2 | 9p | 1.5 | .0399 | .0458 | 2 | 9p | 1.5 | .0248 | .0311 |
| 4 | 3p | .5 | .0632 | .0759 | 4 | 3p | .5 | .0472 | .0639 |
| 4 | 6p | .5 | .0262 | .0477 | 4 | 6p | .5 | .0177 | .0369 |
| 4 | 9p | .5 | .0176 | .0338 | 4 | 9p | .5 | .0109 | .0261 |
| 4 | 3p | 1.0 | .0771 | .0731 | 4 | 3p | 1.0 | .0460 | .0566 |
| 4 | 6p | 1.0 | .0335 | .0433 | 4 | 6p | 1.0 | .0181 | .0301 |
| 4 | 9p | 1.0 | .0191 | .0308 | 4 | 9p | 1.0 | .0107 | .0205 |
| 4 | 3p | 1.5 | .0849 | .0678 | 4 | 3p | 1.5 | .0461 | .0476 |
| 4 | 6p | 1.5 | .0362 | .0400 | 4 | 6p | 1.5 | .0180 | .0260 |
| 4 | 9p | 1.5 | .0221 | .0299 | 4 | 9p | 1.5 | .0122 | .0174 |
| 6 | 3p | .5 | .0431 | .0539 | 6 | 3p | .5 | .0281 | .0448 |
| 6 | 6p | .5 | .0181 | .0356 | 6 | 6p | .5 | .0116 | .0270 |
| 6 | 9p | .5 | .0119 | .0254 | 6 | 9p | .5 | .0068 | .0195 |
| 6 | 3p | 1.0 | .0493 | .0544 | 6 | 3p | 1.0 | .0291 | .0419 |
| 6 | 6p | 1.0 | .0225 | .0329 | 6 | 6p | 1.0 | .0118 | .0216 |
| 6 | 9p | 1.0 | .0133 | .0235 | 6 | 9p | 1.0 | .0073 | .0147 |
| 6 | 3p | 1.5 | .0572 | .0516 | 6 | 3p | 1.5 | .0289 | .0345 |
| 6 | 6p | 1.5 | .0262 | .0319 | 6 | 6p | 1.5 | .0115 | .0193 |
| 6 | 9p | 1.5 | .0151 | .0219 | 6 | 9p | 1.5 | .0084 | .0128 |

Note: each of above values is based on 2,000 replications.

Table 6
Global outlier: Analysis of variance for MISDIF

| Source | DF | Sum of Squares | F value | P | $\eta^2$ * |
|---|---|---|---|---|---|
| NVAR | 2 | 54.9690 | 8700.18 | .0001 | .1388 |
| SIZE | 2 | 79.2299 | 12540.06 | .0001 | .1885 |
| DISTANCE | 2 | .6965 | 110.24 | .0001 | .0020 |
| LOC | 1 | 10.0974 | 3196.33 | .0001 | .0029 |
| NVAR*SIZE | 4 | 27.8604 | 2204.80 | .0001 | .0755 |
| NVAR*DISTANCE | 4 | .0344 | 2.72 | .0277 | .0001 |
| DISTANCE*SIZE | 4 | .2504 | 19.82 | .0001 | .0007 |
| SIZE*LOC | 2 | 2.6439 | 418.46 | .0001 | .0077 |
| NVAR*LOC | 2 | 1.2074 | 191.11 | .0001 | .0035 |
| DISTANCE*LOC | 2 | .6570 | 104.00 | .0001 | .0019 |
| NVAR*DISTANCE *SIZE | 8 | .0683 | 2.70 | .0057 | .0002 |
| NVAR*SIZE*LOC | 4 | .3525 | 27.90 | .0001 | .0010 |
| NVAR*DISTANCE *LOC | 4 | .0645 | 5.11 | .0004 | .0002 |
| DISTANCE*SIZE *LOC | 4 | .2976 | 23.56 | .0001 | .0008 |
| NVAR*DISTANCE *SIZE*LOC | 8 | .0535 | 2.12 | .0308 | .0001 |
| ERROR | 107946 | 341.0093 | | | |
| TOTAL | 107999 | 519.4925 | | | |

* $\eta^2_{effect} = SS_{effect} / \{SS_{effect} + SS_{error}\}$

Table 7
Global outlier: Means and standard deviations of MISDIF for NVAR*SIZE

| Level of NVAR | Level of SIZE | MEAN | SD |
|---|---|---|---|
| 2 | 3p | .1437 | .1148 |
| 2 | 6p | .0503 | .0640 |
| 2 | 9p | .0304 | .0463 |
| 4 | 3p | .0607 | .0667 |
| 4 | 6p | .0250 | .0388 |
| 4 | 9p | .0154 | .0274 |
| 6 | 3p | .0393 | .0487 |
| 6 | 6p | .0169 | .0292 |
| 6 | 9p | .0105 | .0204 |

Note: Each of these rounded values are based on 12,000 replications.

Figure 2
Global outlier: Means of MISDIF by NVAR*SIZE

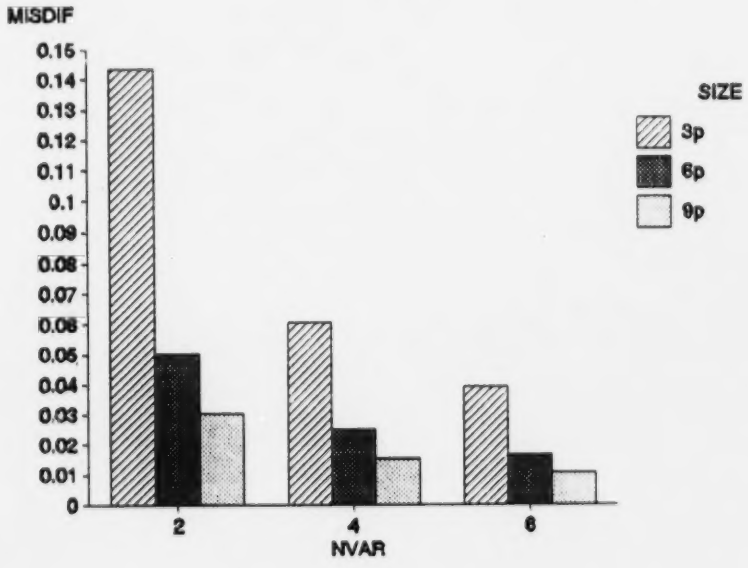

Table 8 Global outlier: t-statistic, P, and $\eta^2$ for levels of NVAR*SIZE for MISDIF

| NVAR*SIZE | 2*3p | 2*6p | 2*9p | 4*3p | 4*6p | 4*9p | 6*3p | 6*6p |
|---|---|---|---|---|---|---|---|---|
| 2*6p | 128.7750<br>.0000<br>@@ .1332 | | | | | | | |
| 2*9p | 156.2060<br>.0000<br>@@ .1844 | 27.4250<br>.0000<br>.0069 | | | | | | |
| 4*3p | 114.3460<br>.0000<br>@@ .1080 | 14.4320<br>.0000<br>.0019 | 41.8585<br>.0000<br>.0160 | | | | | |
| 4*6p | 163.6710<br>.0000<br>@@ .1988 | 34.8970<br>.0000<br>.0112 | 7.4642<br>.0000<br>.0005 | 49.3227<br>.0000<br>.02200 | | | | |
| 4*9p | 176.7900<br>.0000<br>@@ .2245 | 48.0150<br>.0000<br>.0110 | 20.5804<br>.0000<br>.0039 | 62.4496<br>.0000<br>.0349 | 13.1227<br>.0000<br>.0016 | | | |
| 6*3p | 143.9080<br>.0000<br>@@ .1610 | 15.1280<br>.0000<br>.0021 | 12.3010<br>.0000<br>.0014 | 29.5637<br>.0000<br>.0080 | 19.7644<br>.0000<br>.0036 | 32.8888<br>.0000<br>.0099 | | |
| 6*6p | 174.7100<br>.0000<br>@@ .2204 | 45.9330<br>.0000<br>.0192 | 18.5071<br>.0000<br>.0032 | 60.3645<br>.0000<br>.0327 | 11.0397<br>.0000<br>.0011 | 2.1052<br>.4731<br>.0000 | 30.8013<br>.0000<br>.0087 | |
| 6*9p | 183.6490<br>.0000<br>@@ .2381 | 54.8730<br>.0000<br>.0271 | 27.4424<br>.0000<br>.0069 | 69.3067<br>.0000<br>.0426 | 19.9834<br>.0000<br>.0037 | 6.7983<br>.0000<br>.0004 | 39.7404<br>.0000<br>.0144 | 8.9315<br>.0000<br>.0007 |

Note: The values in each cell are: t(5998), P, and $\eta^2$ respectively.
@@ indicates a significant difference.
each t-value is based on 5998 degrees of freedom.

Figure 3
Global outlier: Standard deviations of MISWO and MISWITH

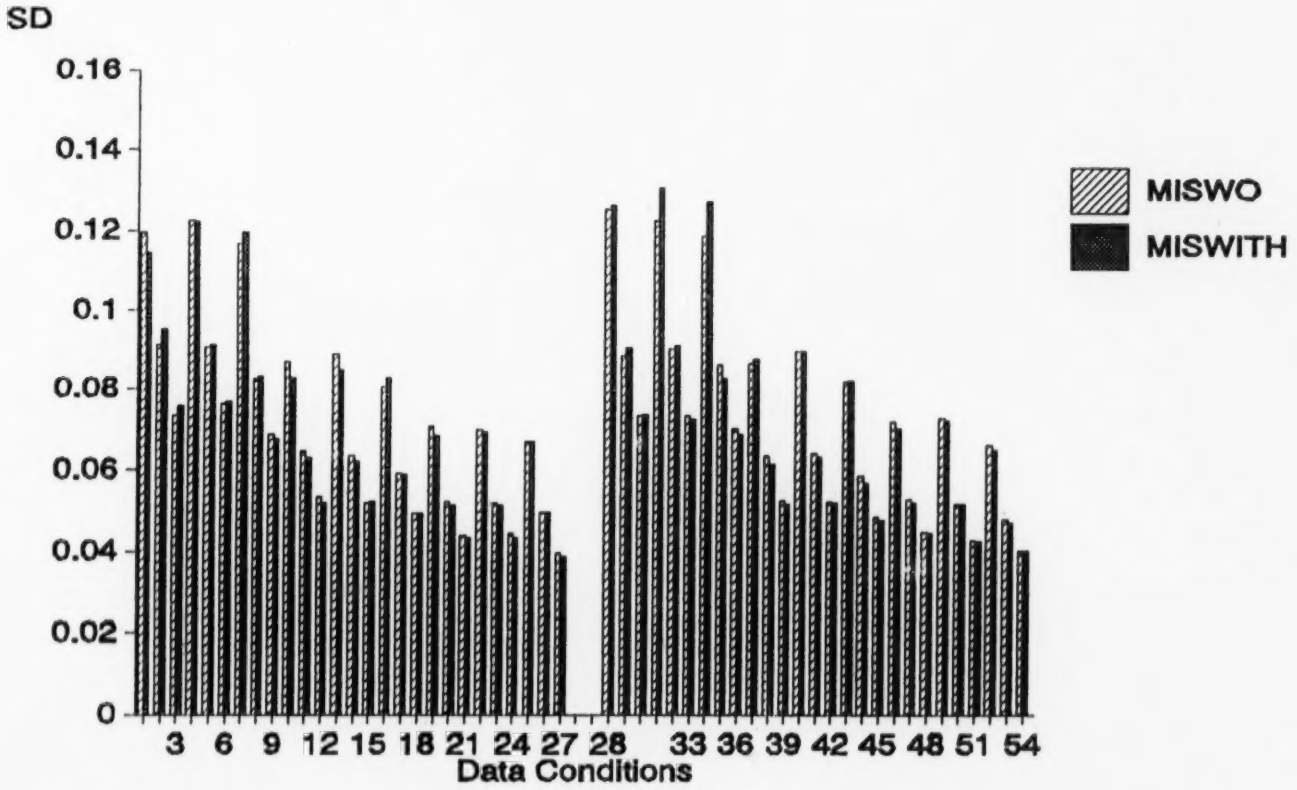

Table 9

Conditions of concern about outlier influence

| DISTANCE | P <.02 | | | .02 ≤ P <.05 | | |
|---|---|---|---|---|---|---|
| | 0.5 | 1.0 | 1.5 | 0.5 | 1.0 | 1.5 |
| N = 12 | 1/2 | 1/2 | 1/2 | 1/3 | 2/3 | 2/3 |
| 12 < N ≤ 24 | 2/4 | 2/4 | 3/4 | 2/- | 4/- | 4/- |
| 24 < N ≤ 36 | 2/- | 3/- | 3/4 | 3/- | 4/- | -/- |
| 36 < N ≤ 48 | 3/- | 3/- | 4/- | 4/- | 4/- | -/- |
| 48 < N ≤ 72 | 3/- | 4/- | 4/- | 4/- | -/- | -/- |
| N > 72 | -/- | -/- | -/- | -/- | -/- | -/- |

Entry definitions:

1: a mean value for PCHNG or MISDIF of greater than .25;
2: a mean value for PCHNG or MISDIF of at least .15 but less .25;
3: a mean value for PCHNG or MISDIF of at least .10 but less .15;
4: a mean value for PCHNG or MISDIF of at least .05 but less .10; and
-: a mean value for PCHNG or MISDIF of less than .05.